Apprentissage non supervisé

Françoise Fessant
TECH/SUSI

28/09/2006

Introduction (1)

- Apprentissage non supervisé
  - Consiste à inférer des connaissances sur les données
    - Sur la seule base des échantillons d'apprentissage,
    - Pas de cible, recherche de structures naturelles dans les données

- Différentes tâches sont associées à l'apprentissage non supervisé
  - Clustering (segmentation, regroupement) : construire des classes automatiquement en fonction des exemples disponibles
    - L'apprentissage non supervisé est très souvent synonyme de clustering
  - Règles d'association : analyser les relations entre les variables ou détecter des associations
  - Réduction de dimensions

Introduction (1)

- Apprentissage non supervisé
  - Consiste à inférer des connaissances sur les données
    - Sur la seule base des échantillons d'apprentissage,
    - Pas de cible, recherche de structures naturelles dans les données

- Différentes tâches sont associées à l'apprentissage non supervisé
  - Clustering (segmentation, regroupement) : construire des classes automatiquement en fonction des exemples disponibles
    - L'apprentissage non supervisé est très souvent synonyme de clustering
  - Règles d'association : analyser les relations entre les variables ou détecter des associations
  - Réduction de dimensions

Introduction (2)

- Quelques bonnes raisons de s'intéresser à l'apprentissage non supervisé
  - Profusion d'enregistrements et de variables
  - Constituer des échantillons d'apprentissage étiquetés peut être très couteux
  - Découvertes sur la structure et la nature des données à travers l'analyse exploratoire
  - Utile pour l'étude de caractéristiques pertinentes
  - Prétraitement avant l'application d'une autre technique de fouille de données

- On obtient des modèles descriptifs qui permettent mieux connaître ses données, de découvrir des informations cachées dans la masse des données

Clustering

- Les techniques de clustering cherchent à décomposer un ensemble d'individus en plusieurs sous ensembles les plus homogènes possibles
  - On ne connaît pas la classe des exemples (nombre, forme, taille)

- Les méthodes sont très nombreuses, typologies généralement employées pour les distinguer
  - Méthodes de partitionnement / Méthodes hiérarchiques
  - Avec recouvrement / sans recouvrement
  - Autre : incrémental / non incrémental

- D'éventuelles informations sur les classes ou d'autres informations sur les données n'ont pas d'influence sur la formation des clusters, seulement sur leur interprétation

Clustering

- Les techniques de clustering cherchent à décomposer un ensemble d'individus en plusieurs sous ensembles les plus homogènes possibles
  - On ne connaît pas la classe des exemples (nombre, forme, taille)

- Les méthodes sont très nombreuses, typologies généralement employées pour les distinguer
  - Méthodes de partitionnement / Méthodes hiérarchiques
  - Avec recouvrement / sans recouvrement
  - Autre : incrémental / non incrémental

- D'éventuelles informations sur les classes ou d'autres informations sur les données n'ont pas d'influence sur la formation des clusters, seulement sur leur interprétation

Similarité

- On cherche à maximiser la similarité intra-classe et à minimiser la similarité inter-classe
- Le point clé est le critère de similarité (fonction de distance)
- La notion de similarité est fonction du type des données
  - Variables définies sur une échelle d'intervalles, binaire, nominale, ordinale, …
- Exemples de distances
  - Distance euclidienne pour les variables continues $d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
  - Coefficient de concordance pour les variables binaires
    - À partir de la table de contingence des variables

		x	
		1	0
y	1	a	b
	0	c	d

$$d(x,y) = \frac{b + c}{a + b + c + d}$$

  - Et bien d'autres mesures possibles

La méthode de clustering "idéale" (1)

- Prise en compte de données de différents types : symbolique, numérique, matrice de similarité, données mixtes
  - Comment comparer des objets caractérisés à la fois par des attributs numériques et symboliques
- Formes arbitraire des clusters
  - De nombreuses méthodes favorisent la découverte de formes sphériques
- Minimisation du nombre de paramètres à fixer
  - Avec certaines méthodes, il est nécessaire de préciser le nombre de classes recherchées
- Insensibilité à l'ordre de présentation des exemples
  - Certaines méthodes ne génèrent pas les mêmes classes si l'ordre de parcours des données est modifié

La méthode de clustering "idéale" (2)

- Résistance au bruit et aux anomalies
  - La recherche des points isolés (outliers) est un sujet de recherche en soi
- Problème en grande dimensions
  - Certaines méthodes ne sont applicables que si les objets ou individus sont décrits sur deux ou trois dimensions
  - Construction et interprétation de clusters en grande dimension
- Passage à l'échelle
  - Certaines méthodes ne sont pas applicables sur de gros volumes de données
- Interprétation et utilisation des résultats
  - Les utilisateurs doivent pouvoir donner un sens aux classes découvertes
  - Comment un utilisateur peut il influencer un processus de catégorisation en introduisant des contraintes

Méthodes de partitionnement

- Construire une partition en classes d'un ensemble d'objets ou d'individus

- Principe
  - Création d'une partition initiale de K classes,
  - Puis itération d'un processus qui optimise le partitionnement en déplaçant les objets d'une classe à l'autre

- Solutions
  - Méthode exacte
    - Par évaluation de toutes les partitions possibles
  - Méthodes approchées/heuristiques
    - K-moyennes (centres mobiles) : une classe est représentée par son centre de gravité
    - K-medoids : une classe est représentée par son élément le plus représentatif

K-moyennes (1)

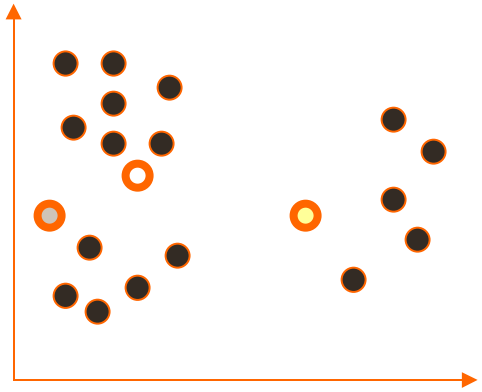
- Principe
  - Une classe est représentée par son centre de gravité,
  - Un objet appartient à la classe dont le centre de gravité lui est le plus proche

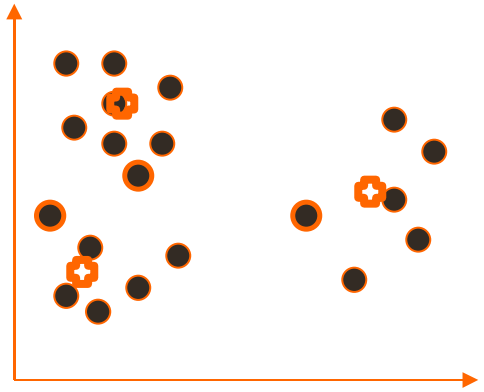
- Algorithme
  - Fixer le nombre de classes a priori K
  - Initialiser aléatoirement les centres (K individus tirés au hasard comme représentants des classes)
  - Affecter chaque observation à la classe qui minimise la distance entre le centre et l'observation (centre le plus proche)
  - Recalculer les nouveaux vecteurs centres (la moyenne)
  - Itérer l'opération jusqu'à la stabilité

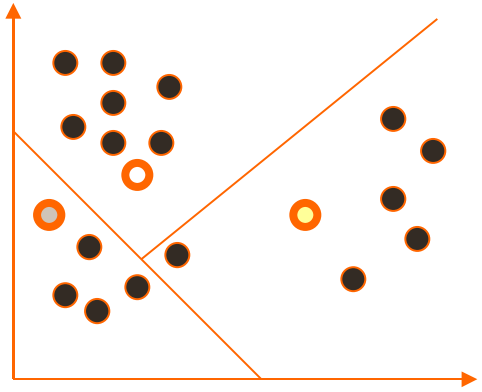
  - Le critère de similarité est l'erreur quadratique entre un exemple et son centre

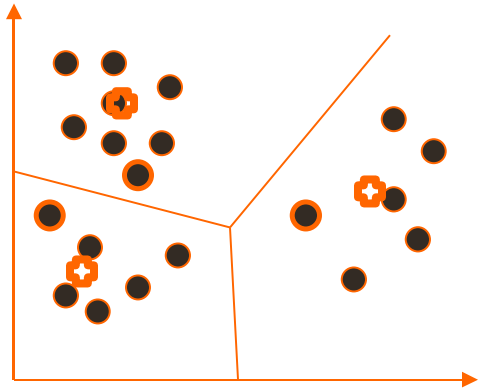
- C'est un algorithme très utilisé

K-moyennes (2)









Choix des centres initiaux au hasard

Affectation des classes

Recalcul des centres

Réaffectation des classes

K-moyennes (3)

- Points forts
  - Simple à implémenter et à comprendre
  - Passage à l'échelle
    - Complexité en O(T.K.N) où T est le nombre d'itérations et N le nombre d'exemples
    - En général K et T sont nettement inférieurs à N

- Difficultés
  - Nécessite de spécifier le nombre de classes à l'avance
  - Utilisable seulement quand la notion de moyenne existe (donc pas sur des données nominales)
  - L'algorithme peut rester bloqué sur un optimum local
  - Sensible au bruit et aux anomalies
  - Sensibilité aux conditions d'initialisation (on n'obtiendra pas forcément les mêmes classes avec des partitionnements initiaux différents)
  - Ne permet pas de définir des classes aux formes non convexes

K-moyennes, variantes

- K-medoids
  - Même principe que les K-moyennes, on utilise à la place de la moyenne le medoid
  - Un medoid est l'objet le plus central dans la classe
  - moins sensible au bruit et aux points isolés, mais plus couteux en calculs
  - Algorithme PAM (Partitioning around medoids)
  - CLARA (Clustering Large Applications), adapté aux grands ensembles d'apprentissage
- Version floue de l'algorithme des K-moyennes
  - Recouvrement possible des classes
- Clustering à base de modèle en utilisant les mixtures de gaussiennes
  - La valeur de K correspond au nombre de composantes de la mixture
  - Apprentissage avec l'algorithme EM : un objet est affecté à une classe selon un poids qui représente sa probabilité d'appartenance à la classe
  - Autorise des formes plus flexibles pour les clusters

Méthodes hiérarchiques (1)

- Les clusters sont organisés sous la forme d'une structure d'arbre

- Deux familles d'approches hiérarchiques
  - Approche ascendante (ou agglomération)
    - Commence avec un objet dans chaque classe,
    - Puis fusionne successivement les 2 classes les plus proches
    - S'arrête quand il n'y a plus qu'une classe contenant tous les exemples

  - Approche descendante (ou par division)
    - Tous les objets sont initialement dans la même classe
    - Puis division de la classe en sous classes jusqu'à ce qu'il y ait suffisamment de niveaux ou que les classes ne contiennent plus qu'un seul exemple

  - Le nombre de clusters n'a pas à être prédéfini

Méthodes hiérarchiques (2)

- Choix d'un critère de distance entre les clusters et d'une stratégie pour l'agrégation/la division

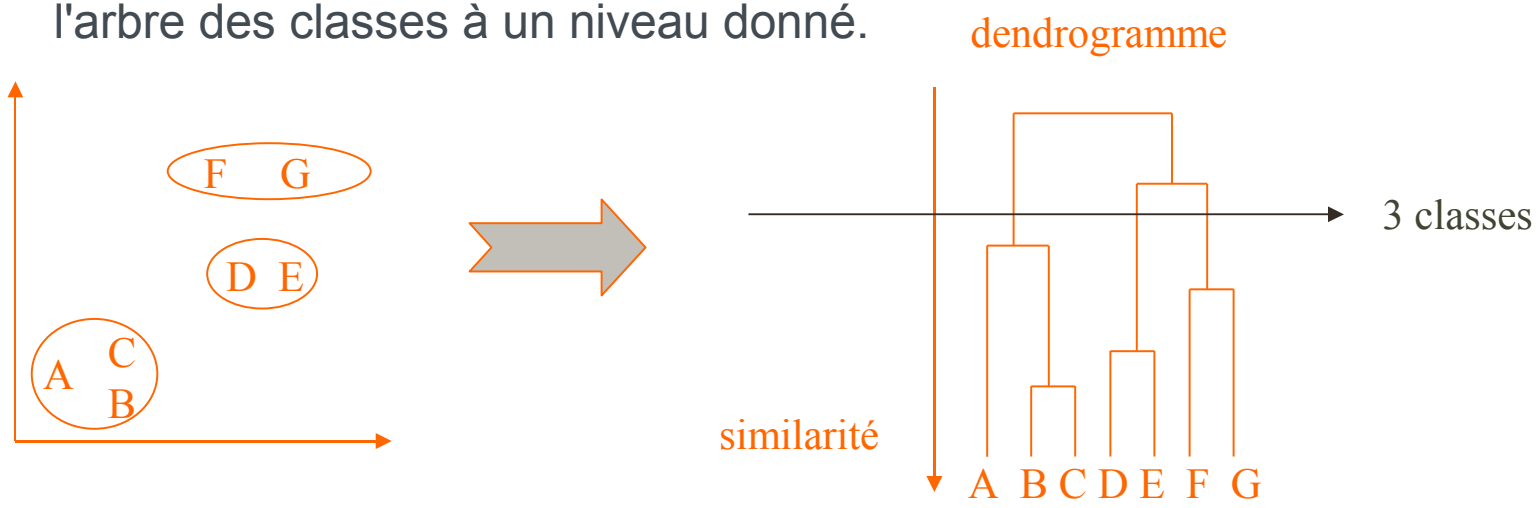
- Les mesures de distance sont nombreuses
  - Lien simple : distance des deux objets les plus proches des clusters,
    - Tendance à former des chaines
  - Lien complet : distance des deux objets les plus éloignés dans les deux clusters,
    - Bonne méthode pour les grappes, mais sensible aux anomalies
  - Lien des moyennes : la distance entre deux objets est donnée par la distance des centres
    - Faible résistance aux anomalies
  - Lien moyen: distance moyenne de toutes les distances entre un membre d'un cluster et un membre de l'autre
    - Plus difficile à calculer

Méthodes hiérarchiques (4)

- Complexité
  - Tous les algorithmes sont au moins en $O(N^2)$ avec N le nombre d'exemples

- Le résultat de la classification hiérarchique est un arbre de classes représentant les relations d'inclusion entre classes (dendrogramme)

- Une classification particulière s'obtient en coupant l'arbre des classes à un niveau donné.



Méthodes hiérarchiques (5)

- Intérêt
  - Facile à implémenter
  - Fournit une structure (préférable pour une analyse détaillée qu'une méthode de partitionnement)

- Difficultés
  - Passage à l'échelle difficile
    - Complexité en $O(N^2)$
  - Pas de remise en question possible des divisions ou agglomérations
    - Deux classes agglomérées à un niveau ne peuvent pas être séparées à un autre niveau

- Recherches en cours
  - Intégration des méthodes hiérarchiques avec d'autres méthodes
    - BIRCH (1996) : basée sur la représentation d'une classe par ses traits caractéristiques
    - CHAMELEON (1999) : basée sur la théorie des graphes et une représentation plus riche des données

Méthodes à base de densité

- Principe de base
  - Utilisation de la densité à la place de la distance
  - Les clusters sont des régions de l'espace qui ont une grande densité de points
    - Un point est dense si le nombre de ses voisins dépasse un certain seuil
    - Un point est voisin d'un autre point s'il est à une distance inférieure à une valeur fixée

- Caractéristiques des méthodes
  - Découvre des classes de formes quelconques
  - Peu sensible à la présence de bruit ou de points isolés
  - Nécessite seulement un parcours des données
  - Ne nécessite pas de prédéfinir un nombre de classes

- Différents algorithmes
  - DBSCAN (1996), OPTICS (1999), DENCLUE (1998), CLIQUE (1998)

Méthodes à base de grille

- Principe
  - Méthode basée sur le découpage de l'espace des exemples suivant une grille
  - Après initialisation, toutes les opérations de clustering sont réalisées sur les cellules, plutôt que sur les données
  - Construction des classes en assemblant les cellules voisines en terme de distance

- Différents algorithmes
  - STING (a STatistical Information Grid approach (1997)
  - WaveCluster (utilise la notion d'ondelettes) (1998)

Conclusion sur le clustering (1)

- Les méthodes sont nombreuses
  - Toutes sont basées sur une mesure de distance ou similarité entre classes
- Différentes classes de méthodes
  - Partitionnement, hiérarchique, …
- Choix du meilleur algorithme ?
  - Chaque algorithme a ses limites et fonctionne pour un certain type de données
  - La qualité du résultat de clustering dépend des connaissances apriori de l'utilisateur, de l'algorithme, de la fonction de distance, de l'application …

Conclusion sur le clustering (2)

- Évaluation d'un résultat de clustering
  - La qualité d'un clustering est difficile à évaluer : les "bons clusters" ne sont pas connus
  - Critères d'évaluation
    - Jugement d'un expert ou évaluation par un utilisateur
    - Utiliser des données étiquetées si elles existent
    - Comparaison avec une segmentation de référence
    - Autres critères : indices divers reposant généralement sur des rapports de distances intra / extra clusters
- Recherches en cours
  - Passage à l'échelle pour traiter de gros volumes de données
  - Prise en compte de contraintes ou connaissances a priori sur les données pour optimiser la construction des classes
  - Méthodes interactives permettant de prendre en compte les besoins des utilisateurs

Découverte de règles d'association (1)

- Identifier les items qui apparaissent souvent ensemble lors d'un évènement (découverte des corrélations entre attributs)

- Règles : Implications du type si X alors Y ou (X->Y)
  - Exemple analyse des ventes : produits fréquemment achetés ensemble {pain, beurre} -> {lait}
  - Les attributs à inclure dans la relation sont inconnus

caddie	liste des items achetés
1	{pain, beurre, lait}
2	{pain, viande}
…	
n	{fruit, poisson,pain}

- Deux critères d'évaluation : support et confiance
  - Support (fréquence) : % d'observations qui contiennent tous les éléments de la règle

Sur 100

30 contiennent à la fois

Support = 30

Découverte de règles d'association (2)

- Deux critères d'évaluation : support et confiance
  - Confiance (précision/fiabilité) : % d'observations présentant Y sachant qu'ils ont X

Sur les 100 33 contiennent parmi ceux-ci 30 contiennent

  - Deux mesures d'intérêt dont le niveau est défini par l'utilisateur

- Seules les règles qui ont un certains support et une certaine valeur de confiance sont intéressantes

- La plupart des approches utilisent des algorithmes basés sur le principe de la détection des ensembles d'items fréquents

Algorithmes (1)

- Algorithme fondateur : APriori (1994), se décompose en 2 étapes
  - Recherche des sous ensembles fréquents (support)
    - Un ensemble fréquent est un ensemble d'items dont la fréquence d'apparition dépasse un certain seuil
  - Recherche des règles d'association (confiance) à partir des sous ensembles retenus

Etape1

3 transactions

n°	items
1	A, B, C
2	A, B, D
3	B, D, E

Taille 1

itemset	support
A	2
B	3
C	1
D	2
E	1

Taille 2

itemset	support
A,B	2
A,D	1
B,D	2

Taille 3

itemset	support
A,B,D	1

Critères : support >40% et Confiance > 70%

Etape2

règles	confiance
A -> B	2/2=100%
B -> A	2/3=66%
B-> D	2/3=66%
D->B	2/2=100%

Algorithmes (2)

- A Priori très simple, mais des limites
  - Le nombre de lectures de la base est important : on commence par lister tous les ensembles de 1 items fréquents, puis tous les ensembles de 2-items fréquents, …
  - Problèmes de mémoire : c'est couteux de générer des ensembles d'items avec beaucoup d'items
- Améliorations : différentes optimisations proposées
  - Génération plus efficace des règles à partir des ensembles d'items
  - Optimisation de la recherche des sous ensembles fréquents
    - FPTree (2004) : détermine les ensembles fréquents sans générer les candidats
    - Représentation condensée de la base sous la forme d'un arbre

Algorithmes (3)

- Différents indices permettent de mesurer la pertinence d'une règle
  - Mesures objectives
    - Support et confiance
  - Mesures subjectives
    - Une règle est intéressante si elle est inattendue et/ou actionnable (si l'utilisateur peut faire quelque chose avec)
  - Le choix d'un indice dépend des points à mettre en évidence et des données
  - Exemple du lift (ou amélioration/ intérêt) : $lift(A \Rightarrow C) = \frac{P(A \cap C)}{P(A)P(C)}$
    - Les transactions contenant A et C sont elles plus nombreuses que si l'achat de A et l'achat de C étaient indépendants
- Extensions
  - Transactions séquentielles
    - Quels sont les produits fréquemment achetés ensembles et successivement ?

Conclusion

- L'apprentissage non supervisé permet d'inférer des connaissances à partir d'échantillons non étiquetés
  - C'est très utile en pratique
- Il existe de nombreuses méthodes et il n'y en a pas une meilleure que les autres
- On choisit la méthode en fonction
  - Du problème et de ses a priori
  - Des données (continues, incomplètes, manquantes, volumineuses, denses)
  - De la tâche à réaliser
  - Du temps de calcul dont on dispose
- De quoi n'a-t-on pas parlé ?
  - Clustering flou, clustering spectral, clustering basé sur les graphes, co-clustering …